Clase 5: Multiplicadores de Lagrange

Ignacio Ramírez

29 de septiembre de 2020

1 Introducción

Consideremos el siguiente problema de optimizacin con restricciones de igualdad:

$$\begin{array}{ll} \min_x & f(x) \\ \text{sujeto a} & h(x) = 0 \end{array} \tag{1}$$

donde $f(x) : \mathbb{R}^n \to \mathbb{R}$ y $h(x) : \mathbb{R}^n \to \mathbb{R}^m$.

La teoría de los *multiplicadores de Lagrange* permite caracterizar las condiciones necesarias y suficientes para que un punto x^* sea mínimo local de (1) bajo ciertas hipótesis adicionales sobre $f(x)$ y $h(x)$. Dichas condiciones se expresan en términos del llamado *lagrangeano*,

$$L(x, \lambda) = f(x) + \lambda h(x). \tag{2}$$

Además de lo anterior, el estudio del lagrangeano (2) permite obtener información adicional sobre la *sensibilidad* del problema original, esto es, cuánto varía $f(x^*)$ en función de variaciones en la restricción $h(x)$. Finalmente, y talvez más importante para este curso, el lagrangeano (2) forma la base de varios de los principales métodos utilizados hoy en día para resolver problemas de optimización complejos.

2 Intuición

De la definición del problema (1) tenemos que x^* pertenece a la curva de nivel 0 de $h(x)$. Por simplicidad, supongamos en adelante que $h(x) \in \mathbb{R}$ (la generalización a $h(x) \in \mathbb{R}^m$ tiene algún tecnicismo adicional pero conceptualmente es lo mismo).

Si $h(x)$ es diferenciable, podemos decir que las *direcciones de variación factibles* d pertenecen al plano *tangente* a dicha curva de nivel, cuya normal es el *gradiente* de $h(x)$. Dicho de otro modo, las direcciones factibles son $\{d : \nabla h(x^*)^T d = 0\}$.

Si $f(x)$ es diferenciable, por ser x^* óptimo local, la variación de $f(x)$ en todas las direcciones factibles d a partir de $f(x^*)$ debe ser nula,

$$f(x^* + d) - f(x^*) = \nabla f(x)^T d + o(d) \to 0.$$

Dicho de otra manera, las direcciones d factibles también deben ser perpendiculares a $\nabla f(x^*)$.

De arriba se concluye que $\nabla f(x^*)$ debe ser *colineal* con $\nabla h(x^*)$, o sea que debe existir λ^* tal que $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$. Esa es la idea general, y esa es la condición necesaria "básica" de esta teoría. A λ^* se le llama el *multiplicador de Lagrange*. La figura 1 intenta mostrar lo anterior gráficamente. Yendo al caso de $h(x) : \mathbb{R}^n \to \mathbb{R}^m$, el detalle técnico necesario para que se pueda generalizar la condición, es que $\nabla f(x^*)$ pueda expresarse como combinación lineal de las m componentes de $h(x)$,

$$\nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) = 0$$

Para que lo anterior se cumpla siempre, el rango del conjunto $\{h_i(x^*) : i = 1, \ldots, m\}$ debe ser m. Cuando eso sucede, x^* es llamado *punto regular*.

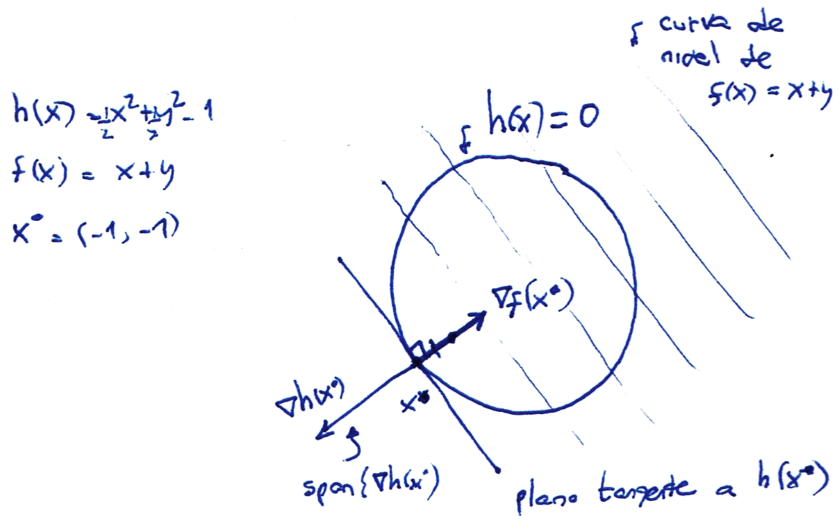

Figura 1: Caso $f(x) = x + y$, $h(x) = (1/2)x^2 + (1/2)y^2 - 1$. El mínimo $x^* = (-1, -1)$ es regular, con $\lambda^* = 1$.

3 El lagrangeano y las condiciones de segundo orden

Observemos ahora que $\nabla f(x^*) + \lambda^* \nabla h(x^*) = \nabla L(x^*, \lambda^*)$. Inmediatamente vemos que (x^*, λ^*) cumple la condición necesaria de mínimo local del problema

$$\begin{array}{ll} \min & L(x, \lambda) \\ \text{sujeto a} & h(x) = 0 \end{array} \tag{3}$$

Haciendo una analogía con las condiciones vistas en la parte anterior del curso, podemos arriesgar que si $L(x, \lambda)$ es doblemente diferenciable, debería cumplirse también que su hessiana, $\nabla^2 L(x, \lambda)$ sea semidefinida positiva en las direcciones factibles, $\{d : \nabla h(x^*)^T d = 0\}$. Como veremos a continuación, ese es precisamente el caso.

4 Enunciado: condiciones necesarias

El siguiente teorema resume y formaliza la discusión anterior:

Teorema 1 (multiplicadores de Lagrange – condiciones necesarias)**.** *Sea x^* un mínimo local de $f(x)$ sujeto a $h(x) = 0$, con $f(x)$, $h(x)$ diferenciables y $\{\nabla h_1(x^*), \nabla h_2(x^*), \ldots, \nabla h_m(x^*)\}$ de rango m. Entonces existe un único vector $\lambda^* \in \mathbb{R}^m$, llamado* vector de multiplicadores de Lagrange*, tal que*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0. \tag{4}$$

Si además, tanto $f(x)$ como $h(x)$ son doblemente diferenciables, entonces se cumple que

$$d^T \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) d \geq 0 \tag{5}$$

para toda variación factible $\{d : \nabla h_i(x^*)^T d = 0, \forall i\}$.

No vamos a demostrar este teorema aquí. La prueba es sencilla pero no aporta mucho a lo que sigue; el estudiante interesado puede leerla en el libro de Bertsekas. Sí vamos a ver la prueba de la suficiencia, que usa técnicas similares pero más útiles en lo que sigue.

4.1 Comentarios

Las condiciones necesarias anteriores pueden escribirse de manera muy compacta en función del lagrangeano,

$$\begin{array}{cc} \nabla_x L(x^*, \lambda^*) = 0 & \nabla_\lambda L(x^*, \lambda^*) = 0, \\ d^T \nabla^2_{xx} L(x^*, \lambda^*)\, d \geq 0 & \forall d : \nabla h_i(x^*)^T d = 0,\, \forall i. \end{array} \tag{6}$$

Es imporante ver que las condiciones de primer orden definen un sistema de ecuaciones de $n + m$ incógnitas que es satisfecho por todo mínimo local x^*. No todo punto que las cumpla es, por otro lado, un mínimo local; bien podría ser un máximo local.

4.2 Ejemplos

Ver Bertsekas 3.1.3 y 3.1.4

5 Condiciones suficientes y el lagrangeano aumentado

Análogamente a las condiciones suficientes vistas para problemas sin restricciones, las condiciones necesarias se transforman en suficientes si la condición de "hessiana semidefinida positiva" se eleva a "definida positiva".

Teorema 2 (Condiciones suficientes de segundo orden)**.** *El punto x^* es mínimo local estricto de $f(\cdot)$ si se cumple lo siguiente para un cierto multiplicador λ^*:*

$$\begin{array}{cc} \nabla_x L(x^*, \lambda^*) = 0 & \nabla_\lambda L(x^*, \lambda^*) = 0, \\ d^T \nabla^2_{xx} L(x^*, \lambda^*)\, d > 0 & \forall d : \nabla h_i(x^*)^T d = 0,\, \forall i. \end{array} \tag{7}$$

Notar que la única diferencia entre (6) y (7) es el signo de mayor estricto en la segunda linea! Además, en el caso de la suficiencia no es necesario exigir regularidad.

Para probar lo anterior vamos a utilizar una función auxiliar llamada el *lagrangeano aumentado*, que es el resultado de agregarle un término adicional al lagrangeano clásico (2),

$$L_c(x, \lambda) = f(x) + \lambda h(x) + \frac{\mathbf{c}}{\mathbf{2}} \|\mathbf{h}(\mathbf{x})\|^{\mathbf{2}}, \tag{8}$$

con $c > 0$. Notar que (8) es el lagrangeano del siguiente problema,

$$\begin{array}{ll} \min & f(x) + \frac{c}{2}\|h(x)\|^2 \\ \text{sujeto a} & h(x) = 0. \end{array} \tag{9}$$

Claramente, dicho problema tiene los mismos mínimos locales que el problema original. La idea ahora es estudiar las condiciones suficientes de este problema, compararlas con las del original, y ver que son las mismas a partir de un cierto valor de c.

$$\nabla_x L_c(x, \lambda) = \nabla f(x) + (\lambda + c h(x)) \nabla h(x), \tag{10}$$

$$\nabla^2_{xx} L_c(x, \lambda) = \nabla^2 f(x) + \sum_{i=1}^{m} (\lambda_i + c h_i(x)) \nabla^2 h_i(x) + c \nabla h(x) \nabla h(x)^T \tag{11}$$

Si x^* y λ^* cumplen las condiciones (7), tenemos que

$$\nabla_x L_c(x^*,\lambda^*) = \nabla f(x^*) + \nabla h(x^*)[\,\lambda^* + c\,\overbrace{h(x^*)}^{=0}\,] = \nabla f(x^*) + \nabla h(x) = \nabla_x L(x^*,\lambda^*) = 0, \tag{12}$$

$$\nabla^2_{xx} L_c(x^*,\lambda^*) = \nabla^2 f(x^*) + \sum_{i=1}^{m} [\,\lambda_i^* + c\,\overbrace{h_i(x^*)}^{=0}\,] \nabla^2 h_i(x^*) + c\nabla h(x^*)\nabla h(x^*)^T \tag{13}$$

$$= \overbrace{\nabla^2 f(x^*,\lambda^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*)}^{\nabla^2_{xx}L(x^*,\lambda^*)} + c\nabla h(x^*)\nabla h(x^*)^T \tag{14}$$

$$= \nabla^2 L_{xx}(x^*,\lambda^*) + c\nabla h(x^*)\nabla h(x^*)^T \tag{15}$$

Por las hipótesis de suficiencia de segundo orden tenemos además que

$$d^T \nabla^2_{xx} L(x^*,\lambda^*)\, d > 0 \quad , \forall d : \nabla h(x^*)^T d = 0,$$

o lo que es lo mismo,

$$d^T \nabla^2_{xx} L(x^*,\lambda^*)\, d > 0 \quad , \forall d : \nabla h(x^*) \nabla h(x^*)^T d = 0,$$

o sea que toda dirección factible d pertenece al *nucleo* de la matriz *semidefinida positiva*

$$Q = \nabla h(x) \nabla h(x)^T.$$

Más aún, restringida a ese nucleo, la matriz *simétrica* $P = \nabla^2_{xx} L(x^*,\lambda^*)$ es *definida positiva.*

Resulta que para P y Q cumpliendo esas condiciones de "positividad complementaria", existe un $\hat{c} > 0$ tal que para todo $c > \hat{c}$, $P + cQ$ es *definida positiva.*[1] Aplicando lo anterior a (15) llegamos a que

$$\nabla^2_{xx} L_c(x^*,\lambda^*) = \nabla^2 L_{xx}(x^*,\lambda^*) + c\nabla h(x^*)\nabla h(x^*)^T > 0 \quad , \forall\, c > \hat{c}.$$

O sea que para $c > \hat{c}$, x^* es un mínimo local *estricto* del problema sin restricciones $\min L_c(\cdot,\lambda^*)$. De ahí se deriva en particular que

$$L_c(x,\lambda^*) \geq L_c(x^*,\lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2, \quad \forall\, x : \|x - x^*\| < \epsilon, \tag{16}$$

para algunos $\gamma > 0$ y $\epsilon > 0$. Tenemos ahora que $L_c(x^*,\lambda^*) = f(x^*)$ por definición. Si además nos restringimos a $h(x) = 0$, tenemos que $L_c(x,\lambda^*) = f(x)$. Sustituyendo en (16), y aplicando la restricción adicional,

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x - x^*\|^2, \quad \forall\, x : h(x) = 0\,,\ \|x - x^*\| < \epsilon.$$

Entonces x^* es un mínimo local estricto de $f(x)$ sobre $h(x) = 0$. □

5.1 Comentarios sobre el lagrangeano aumentado

En lo que respecta a los objetivos de este curso, una de las cosas más importantes de este tema es lo que en la demostración anterior aparece como una función auxiliar, el *lagrangeano aumentado.*

De la demostración del teorema surge que uno puede obtener el óptimo de un problema con restricciones a partir de la solución de otro sin restricciones. Ésto, por lo menos a primera vista, depende sin embargo de conocer los multiplicadores λ^*, algo que por ahora no sabemos cómo hacer.

Veremos al final de este tema que en la práctica es posible (y se puede demostrar) llegar al óptimo, y a los multiplicadores correspondientes, basándose en aproximaciones de λ^*. Incluso es posible alternar entre aproximaciones de x^* y λ^*. Todo esto cae dentro de la familia de *métodos de multiplicadores de Lagrange*, llegando al ADMM (Alternate Directions Method of Multipliers), muy popular en los últimos años, y que tendrá su apartado especial en este curso.

[1] La interpretación geométrica es bastante intuitiva si uno piensa en la forma cuadrática asociada: P es estrictamente creciente (positiva) para los lados que Q se anula (su nucleo), y a partir de cierto valor de c, la positividad de Q le gana a la negatividad de P, entonces $P + cQ$ pasa a ser definida positiva "para todos lados".

6 Sensibilidad

En muchos problemas, los multiplicadores de Lagrange λ tienen una interpretación interesante (precios en economía, magnitudes físicas). Pero una interpretación bastante general es que dan una idea de *cuánto cambiará el costo óptimo en función de variaciones en las restricciones.* Esto es fácil de ver para un problema con restricciones lineales. Consideremos el problema

$$\begin{array}{ll} \min & f(x) \\ \text{sujeto a} & a^T x = b \end{array} \tag{17}$$

Tenemos $h(x) = a^T x - b$, o sea que $\nabla h(x) = a$. Las condiciones de optimalidad de Lagrange nos dicen entonces que en el óptimo x^* tenemos λ^* tal que $\nabla f(x) + \lambda^* a = 0$ o bien

$$\nabla f(x) = -\lambda^* a. \tag{18}$$

Utilizando (18) en la aproximación de primer orden del costo $f(x^*)$,

$$\Delta\text{costo} = f(x^* + \Delta x) - f(x^*) = \nabla f(x^*)^T \Delta x + o(\|\Delta x\|) = -\lambda^* a^T \Delta x + o(\|\Delta x\|)$$

Por otro lado, por ser una restricción lineal tenemos que $a^T \Delta x = \Delta b$, por lo que

$$\Delta\text{costo} = -\lambda^* \Delta b + o(\|\Delta x\|).$$

Concluimos que la variación de primer orden del costo óptimo al perturbar b es precisamente λ^*.

El teorema general es un poco complicado de enunciar (ver Bertsekas p. 309). La esencia es la siguiente:

Teorema 3. *Sean x^* y λ^* el óptimo y su multiplicador de Lagrange para el problema*

$$\begin{array}{ll} \min & f(x) \\ \text{sujeto a} & h(x) = u \end{array} \tag{19}$$

Sea $p(u)$ el costo óptimo del problema anterior al variar u con $x(u)$ y $\lambda(u)$ mínimo y multiplicador asociados. Entonces

$$\nabla p(u) = -\lambda(u).$$

7 Condiciones de desigualdad

Consideremos ahora el caso en que algunas de las restricciones del problema son de la forma $g_j(x) \leq 0$,

$$\begin{array}{ll} \min & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \ldots, m \\ & g_j(x) \leq 0,\ j = 1, \ldots, r \end{array} \tag{20}$$

Una forma de ver este problema es reducirlo al caso de igualdad de la siguiente manera.

Definamos el *conjunto activo* $A(x) = \{j : g_j(x) = 0\}$, es decir $A(x)$ incluye los índices j de aquellas restricciones $g_j(x)$ que están *activas.*

Supongamos ahora que x^* es un mínimo local del problema (20). Resulta obvio que x^* es también mínimo local de un problema para el cual las restricciones no activas $j \notin A(x^*)$ son descartadas. Por otro lado, las restricciones *activas* en x^* pueden ser tratadas como restricciones de igualdad. Entonces, x^* cumple las condiciones necesarias del teorema de Lagrange,[2]

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla g_j(x^*) = 0.$$

[2] Sujeto al tecnicismo de que x^* sea *regular* para ese problema, es decir, que $\{\nabla h_i(x^*)\} \bigcup \{\nabla g_j(x^*) : j \in A(x^*)\}$ sea un conjunto l.i. Este tecnicismo se puede evitar, como veremos más adelante, pero es necesario para el teorema clásico, las KKT.

Si ahora definimos $\mu_j = 0$ para $j \notin A(x^*)$ llegamos a

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0, \tag{21}$$

$$\mu_j^* = 0, \quad j \notin A(x^*) \tag{22}$$

Las condiciones anteriores, junto a alguna cosita más, constituyen las llamadas condiciones Karush-Kuhn-Tucker (KKT). Un dato adicional de éstas condiciones es que los μ_j^* son no negativos. Esto es fácil de ver siguiendo el argumento de sensibilidad visto anteriormente (pendiente:dibujo). El enunciado formal es el siguiente.

Teorema 4 (Condiciones necesarias de Karush-Kuhn-Tucker)**.** *Sea x^* mínimo local del problema*

$$\begin{array}{ll} \min & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \ldots, m \\ & g_j(x) \leq 0,\ j = 1, \ldots, r, \end{array} \tag{23}$$

donde $f(x)$, $h_i(x)$ y $g_j(x)$ son continuamente diferenciables y x^* es regular. Entonces existen vectores multiplicadores de Lagrange λ^* y μ^* tales que

$$\nabla_x L(x, \lambda^*, \mu^*) = 0, \tag{24}$$

$$\mu_j^* \geq 0, \quad j = 1, \ldots, r \tag{25}$$

$$\mu_j^* = 0, \quad j \in A(x^*), \tag{26}$$

donde $A(x^*)$ es el conjunto de restricciones activas de x^*. Si además $f(x)$, $h_i(x)$ y $g_j(x)$ son doblemente diferenciables entonces

$$d^T \nabla_{xx}^2 L(x, \lambda^*, \mu^*) d \geq 0$$

para $\{d : \nabla h_i(x^*)^T d = 0\,, \nabla g_j(x^*) = 0,\ \forall\, j \in A(x^*)\}$.

No vamos a demostrarlo aquí, si bien la demostración no es difícil (ver Bertsekas p. 316). Cabe notar que todo se extiende naturalmente a este caso, incluso el concepto de que la hessiana de $f(x)$ sea semidefinida positiva en el "subespacio de direcciones factibles", ahora incluyendo las restricciones de desigualdad activas.

También, al igual que en los otros casos, se obtienen las condiciones suficientes (para funciones doblemente diferenciables) al exigir que $\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)$ sea definida positiva sobre el subespacio de direcciones factibles; en este caso se agrega la condición adicional de que $\mu_j^* > 0$ para las restricciones activas (de nuevo evidenciando el crecimiento estricto de la función al momento de llegar al borde del conjunto factible).

8 Condiciones de Fritz John

Todas las condiciones necesarias que vimos anteriormente (no las suficientes) exigen que x^* sea *regular*, algo que en principio no es del todo intuitivo. Para el caso de restricciones de igualdad esto es exigir que los $\{\nabla h_i\}$ sean l.i., de modo que siempre pueda haber una combinación lineal que produzca $-\nabla f$:

$$\nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) = 0.$$

Las condiciones de Fritz John son una versión más moderna de las KKT, y por ende del teorema de Lagrange, que resuelven el problema agregando un multiplicador a $\nabla f(x^*)$,

$$\mu_0^* \nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) + \sum_j \mu_j^* g_j(x^*) = 0,$$

con el requerimiento adicional de que $\mu_0^*, \lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*$ no son todos cero a la vez.